

# Spline regression for zero-inflated models

T. Opitz, P. Tramini, N. Molinari

April 12, 2013

## Abstract

We propose a regression model for count data when the classical GLM approach is too rigid due to a high outcome of zero counts and a nonlinear influence of continuous covariates. Zero-Inflation is applied to take into account the presence of excess zeros with separate link functions for the zero and the nonzero component. Nonlinearity in covariates is captured by spline functions based on B-splines. Our algorithm relies on maximum-likelihood estimation and allows for adaptive box-constrained knots, thus improving the goodness of the spline fit and allowing for detection of sensitivity changepoints. A simulation study substantiates the numerical stability of the algorithm to infer such models. The AIC criterion is shown to serve well for model selection, in particular if non-linearities are weak such that BIC tends to overly simplistic models. We fit the introduced models to real data of children's dental sanity, linking caries counts with the so-called Body-Mass-Index (BMI) and other socio-economic factors. This reveals a puzzling non-monotonic influence of BMI on caries counts which is yet to be explained by clinical experts.

**Keywords:** B-splines; count data; DMFS index; nonlinear regression; overdispersion; Zero Inflation.

## 1 Introduction

Classical log-linear regression for Poisson count data results in an inadequate fit to data if at least one of the two basic assumptions is violated: First, even conditionally on the covariates, the counted events may not arise independently and with identical distribution. In real data, this typically leads to overdispersion, breaking the mean-variance equality of Poisson counts. Second, the assumption of a linear influence of covariates on the log-transformed expectation may be too rigid. In fact, for some applications the detection of changes in sensitivity with respect to a covariate is fundamental.

In this work, we tackle a particular case of the first problem where overdispersion is (partly) due to an excess of zero counts. We assume a binary random effect which determines affiliation of observations to a class of structural zeros or alternatively to the class of ordinary counts. Therefore, we use Zero Inflation (ZI) which proposes a mixture of a Dirac mass in 0 with the respective count

distribution, see [12]. Switching from Poisson to a negative binomial distribution can take account of additional overdispersion in the ordinary component. Applications of ZI regression are numerous in the medical context, e.g. in public health studies like [1] and [2]. Other, more uncommon applications are e.g. software fault prediction in [11] and patent outsourcing rates in [3].

To deal with the second problem of non-linearities, we resort to B-splines (cf. [4]). Spline-based approaches have become common for regression models in statistics, see the presentation [5] on spline smoothing and non-parametric regression, and the Multivariate Adaptive Regression Spline technique (*MARS*) used for normal residuals by [7]. Whereas *smoothing spline* methods use the set of data points as spline knots and impose a smoothness penalty, *regression spline* methods tend to favor a small number of spline knots. For the latter methods, variable knots may be free knots of parametric nature or knots that are chosen by an adaptive method. The smoothness penalty introduced in the intermediate form of *penalized regression splines* allows to use a higher number of knots without overfitting. In summary, splines offer a semi-parametric setting such that the effective model dimension remains controllable

Our approach replaces a linear regression coefficient by a spline curve determined by its degree of curvature, its knots and the coefficients of the resulting B-spline basis. A variant of the Evolutive Boundary Knots approach (*EBOK*, [13]) is applied in order to use adaptive knots in the numerical calculation of the maximum likelihood estimator.

The introduced models for count data regression unite interpretability of parameters and high flexibility with respect to non-standard behavior, but avoid numerical complications by constraining knots. Moreover, iterating the EBOK method with different initial constraints leads asymptotically to globally optimal knots (see [13]). This class of models can be classified among the large class of Generalized Additive Models (GAM; [9]).

In the remainder of the paper, we first present details of the modelisation in the following section. Issues and the algorithm related to the spline modeling approach are discussed in Section 3, followed by remarks on model selection criteria in Section 4. The simulation study in Section 5 sheds light on model choice and goodness-of-fit issues, with particular focus on the selection of a spline space. We apply the presented models to caries count data of 12-year old French children in Section 6, with the Body-Mass-Index as potentially non-linear continuous covariate. Open issues and further possible developments are subject of Section 7.

## 2 The Model

The Zero-Inflated Poisson Process  $\text{ZIP}(\mu, \pi)$  with parameters  $\mu > 0, \pi \in [0, 1)$  has probability mass function (pmf)

$$\begin{cases} \pi + (1 - \pi) \exp(-\mu) & k = 0 \\ (1 - \pi) \frac{\mu^k}{k!} \exp(-\mu) & k = 1, 2, \dots \end{cases}$$

The Zero-Inflated Negative Binomial Process  $\text{ZINB}(\mu, \nu, \pi)$  with expectation  $\mu > 0$ , dispersion parameter  $\nu > 0$  and  $\pi \in [0, 1)$  has pmf

$$\begin{cases} \pi + (1 - \pi) \left( \frac{\nu}{\mu + \nu} \right)^\nu & k = 0 \\ (1 - \pi) \frac{\Gamma(k + \nu)}{k! \Gamma(\nu)} \frac{\nu^\nu \mu^k}{(\mu + \nu)^{k + \nu}} & k = 1, 2, \dots \end{cases}$$

We denote univariate B-splines by  $N_{i,d}(\cdot, \Delta)$ ,  $i = 1, \dots, m + d + 1$ , with a grid  $\Delta$  of knots on a finite interval  $[a, b]$ , boundary knots  $a, b$  and the degree  $d$  of spline smoothness. The degree  $d = 1$  corresponds to linear splines and  $d = 3$  to cubic splines.

We follow Lambert's model of two regression equations as presented in [12], allowing covariates to influence on the probability of a structural zero and on the count expectation. The model for  $n$  observations  $Y_i$  is

$$\begin{cases} Y_i \mid (\mathbf{X}_i, \mathbf{Z}_i) = (\mathbf{x}_i, \mathbf{z}_i) \sim (\mu(\mathbf{x}_i), \pi(\mathbf{z}_i)) & \text{independent} \\ \log \mu(\mathbf{x}_i) = g^c(\mathbf{x}_i, \beta^c) \\ \text{logit } \pi(\mathbf{z}_i) = g^z(\mathbf{z}_i, \beta^z) \end{cases}$$

The covariate vectors  $\mathbf{x}_i, \mathbf{z}_i$  may coincide, e.g. if each component depends on all available covariates. Dependence on covariates can be modeled either by a linear regression coefficient or by a spline function for continuous covariates. I.e.,

$$g^c(\mathbf{x}_i, \beta^c) = \sum_{j=1}^{m^c} g_j^c(x_{ij}, \beta_j^c), \quad g^z(\mathbf{z}_i, \beta^z) = \sum_{j=1}^{m^z} g_j^z(z_{ij}, \beta_j^z)$$

where for  $ind \in \{c, z\}$  either

$$g_j^{ind}(u, \beta_j) = u \beta_j$$

or

$$g_j^{ind}(u, \beta_j) = g_j^{ind}(u, \beta_j, \Delta_j) = \sum_l \beta_j^{(l)} N_{l,d}(u, \Delta_j).$$

If splines are applied for  $m > 1$  covariates  $x_{\cdot,j}, j \in J$ , we remove  $m - 1$  of the terms  $\beta_j^{(1)} N_{1,d}(u, \Delta_j)$ ,  $j \in J$ , to ensure an identifiable model: The "partition of 1"-property  $\sum_l N_{l,d} \equiv 1$  gives

$$\sum_l \beta_j^{(l)} N_{l,d}(u, \Delta_j) = \beta_j^{(1)} + \sum_{l>1} (\beta_j^{(l)} - \beta_j^{(1)}) N_{l,d}(u, \Delta_j),$$

so we combine the  $m$  constants  $\beta_j^{(1)}$  into one single coefficient  $\tilde{\beta}_{j_0}^{(1)} = \beta_{j_0}^{(1)}$ , and for the other coefficients we replace  $\beta_j^{(l)}$  by  $\tilde{\beta}_j^{(l)} = \beta_j^{(l)} - \beta_j^{(1)}$ . If we apply the linear modelisation instead of a spline curve in one of the components, as usual a regression constant can be added.

For cubic splines, we can optionally impose that second-order derivatives of a spline curve are zero at the boundary knots  $a, b$  of the respective covariate's

domain – resulting in a *natural cubic spline*. This counteracts the strong near-boundary variability of non-linear spline curves and reduces the number of free B-spline coefficients by 2. The spline model is flexible in the sense that other algorithmically tractable constraints on the form of the spline curve could be imposed.

The (semi-)parametric nature of the model with coefficients and knots results in models that are flexible and remain conveniently interpretable and comparable, see also the remarks on model selection later on. Quantities like derivatives are analytically accessible.

We fit this model by maximum-likelihood estimation of the  $\beta$ -coefficients, of knot locations  $\Delta$  if considered as variable and of additional parameters that are constant for all observations, e.g. the extent of overdispersion  $\nu$  of ZINB. Other links than the proposed ones may be chosen if convenient. The proposed model is of additive type, but extensions with more refined multivariate spline techniques can deal with interaction effects of covariates.

### 3 Optimal Knot Locations

In general, the set-up of fixed knots is an arbitrary restriction of the set of available spline curves. In particular, the B-spline basis may contain elements that merely improve the fit and model dimension is blown up by insignificant parameters. For instance, think of an equidistant grid of knots which performs poorly if there are large regions of sparse observations in the covariate domain. However, optimizing knot locations makes numerical optimization more challenging for the following reasons:

- The number of free parameters in “good” models tends to increase.
- Analytic partial derivatives with respect to knots are difficult to establish; hence we use numerical derivatives.
- Knots influence differently on the likelihood in comparison with  $\beta$ -coefficients.
- In particular, many suboptimal stationary points which correspond to coinciding knots may arise (see [10] for this *lethargy* property), rendering non-constrained knot optimization infeasible.

We use an Evolutive Bounded Optimal Knots algorithm (see [13]) that imposes iteratively adaptable interval constraints on variable knots, based on a partition of a covariate domain  $[a, b]$ . In general, the resulting optimized knots are not globally optimal, but iterating the procedure with different initial interval constraints would lead to convergence to the optimal set of knots. To avoid coinciding knots and the associated discontinuities in spline functions, a minimal positive distance between two knots can be defined. In the following, we present and apply the algorithm for only one spline-modeled covariate; the extension to an additive model as introduced in Section 2 is straightforward. However, for the inclusion of interaction terms in other multivariate spline approaches, more

complex adaptation steps become an issue. We further exclude iteration with different initial knot constraints and propose instead initial knots corresponding to equiprobable quantiles such that segments between knots contain the same number of observations. We shortly outline the functioning of the algorithm:

- Initial knots are determined according to equiprobable quantiles of the covariate.
- Each knot can vary within the boundaries of an interval, also called its box. Boxes for different knots do not overlap, and usually we separate boxes by a small distance to avoid exactly coinciding knots. Initial boundaries are fixed to the center of the segment between two initial knots.
- We determine initial B-spline coefficients and other estimated parameters by numerical maximisation of the likelihood with the fixed initial knots.
- We maximize likelihood numerically with respect to box-constrained knots with the initial parameter estimates as starting point.
- If one of the knots coincides with a boundary of its box, we shift this boundary such that it is in the middle between this knot and its neighboring knot. The adjacent boundary of the box of the neighboring knot is shifted accordingly such that boxes do not overlap and the separating distance is preserved.
- If one of the knots coincides with a boundary, but the neighboring knot is too close to shift this boundary, we do nothing. Usually this means that there is a sharp break in the curvature of the fitted spline curve.
- If box boundaries have been shifted, we iterate the procedure with the adapted box configuration.
- If no boundaries are to be shifted, we stop the iteration, and the last set of knots and estimated parameters defines the fitted model.

## 4 Model selection

We consider AIC and BIC for model selection. Moreover, (cross-validation) residuals allow us to evaluate a model's predictive power. Difficulty lies in determining the dimension of spline models. With a parametric interpretation of spline coefficients and knots, we count each estimated scalar as a model parameter. So for each spline curve included in the model, we add to the model dimension the sum  $m + d + 1 + m_f$  of the number  $m$  of inner knots, of the spline order  $d + 1$  (2 for linear splines, 4 for cubic splines) and of the number  $m_f \leq m$  of free inner knots. This model dimension corresponds to  $d + 1$  initial parameters associated to the left endpoint  $a$  of the covariate interval  $[a, b]$ , to  $m$  change-points of which  $m_f$  are free parameters, and to the  $m$  respective changes in the  $(d - 1)$ th derivative. For instance, a linear free-knot spline  $s$  with

$d = 1$  possesses an initial level  $s(a)$ , an initial sensitivity  $s'(a)$  and  $m = m_f$  breakpoints with the respective sensitivity changes. If natural cubic splines are applied, the model dimension is reduced by 2.

Spline spaces include all linear functions, but the penalisation of higher model dimension would favor the simpler models with a linear coefficient when non-linear effects are not strong. If the sample is large, BIC penalizes strongly in the semi-parametric setting of splines which may include a rather high number of parameters compared to purely parametric models. This seems disproportionate such that we prefer AIC, a choice which is substantiated through the simulation study in the following section. In general, one should be cautious when comparing models with different nature of the respective parameters. In the literature, no consensus is yet found on how to assess the effective model dimension when splines are involved. Further model checks can be based on a more profound analysis of residuals.

In the application to dental sanity data we opted for a preselection of the models performing best in terms of AIC among a wide variety of available models. To obtain our designated "winner model" and to elude selection bias, we checked these models' predictive power by minimizing a cross-validation mean residual error.

## 5 A simulation study

We focus on two issues:

- Study (1): Retrieve the B-spline coefficients of a simulated model when the spline space is known.
- Study (2): Find the best-fitting spline model when the simulated model is not equivalent to a spline model.

In each case, we use ZIP regression. We assess the (relative) goodness-of-fit for different models based on AIC, BIC and cross-validation mean residual error with respect to raw residuals (MRE). We randomly choose 20 non-overlapping subsamples of equal size for cross-validation to avoid the costly leave-one-out procedure. Moreover, we apply the supremum- and  $L^1$ -norms to measure the deviation of fitted spline curves from the simulated ones.

In Study (1), we restrict the analysis to the more intricate cubic splines which usually lead to stronger variability in fitted curves than linear splines, particularly near the boundaries. We generate 100 simulations of the model

$$X_i \sim U[0, 1], Y_i | X_i = x_i \sim \text{ZIP}(\log^{-1} f_c(x_i), \text{logit}^{-1} f_z(x_i)), \quad (1)$$

with  $i = 1, \dots, 200$ , where  $f_c$  is a cubic spline

$$f_c(x) = \sum_{i=1}^6 \alpha \beta_i N_{i,k}(x)$$

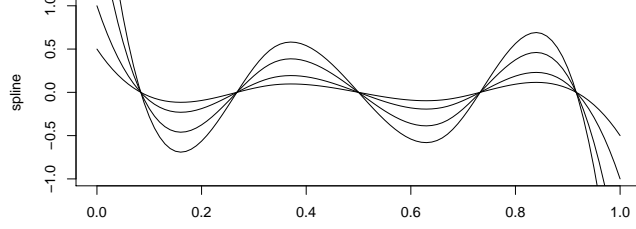


Figure 1: Spline curves for simulated models of Study (1).

with marginal knots 0 and 1 and with 2 equidistant interior knots  $1/3$  and  $2/3$ , and

$$f_z(x) = \beta_0^z + \beta_1^z x = 1 - x .$$

The coefficients  $\alpha\beta_i$  are chosen such that

$$\beta_i = -(-1)^i \quad (i = 1, \dots, 6) ,$$

and  $\alpha \in \{0.5, 1, 2, 3\}$  determines how strongly the spline function oscillates, see Figure 1. Tables 1 to 4 summarize for the four given values of  $\alpha$  the goodness-of-fit statistics for linear fits and cubic spline fits with up to three equidistant fixed knots in the count component. We show median values and standard deviations over the 100 simulations for the proposed criteria (note that the standard deviation coincides for AIC and BIC). Additionally, for AIC, BIC and MRE, the number of times a model performed best is given. Median values and standard deviations for the estimated regression constants and coefficients of the zero component are also included.

The case  $\alpha = 0.5$  with the weakest oscillation is best fitted with a linear model. Even the norm criteria tend to favor the linear model. In general, curves that are "close" to linear are better fitted by linear models, since higher model dimension and increased variability make the spine fit less appealing. With  $\alpha$  increasing, the actual model is better reconstructed.

AIC, MRE and the  $L^1$ -norm show similar preferences, whereas BIC has a stronger preference for the simpler linear model. The fitted linear coefficients of the zero component show some bias in some cases, but no other anomalies.

In Study (2), we keep the model from Study (1) with 50 simulations, where we change  $f_c$  to the non-spline function

$$f_c(x) = 1 + \sqrt{x} + \sin(4\pi x), \quad x \in [0, 1] ,$$

which describes a periodic phenomenon  $\sin(4\pi x)$  with a non-linear trend  $1 + \sqrt{x}$ .

We fit linear splines and cubic splines with fixed and adaptive knots (EBOK method). Tables 5 to 8 show the results. MRE preference is scattered over the range of models such that there is no clear preference. We always find a good correspondance between AIC with  $L^1$ -norm and partly with the supremum norm, so AIC serves well in finding the best model. BIC tends to simpler models. Variable knots tend to improve the examined criteria, and less knots are needed to obtain the optimal value of a criterion. However, an improvement in best attainable BIC values seems questionable when switching to variable knots.

We conclude that, in practice, the presented class of ZI spline regression models allows for fitting to data with the proposed fitting algorithm which is sufficiently robust. BIC works well if non-linearities are strong, in other cases we advise AIC, which can be complemented by the analysis of residuals, in particular of MRE.

## 6 Application to dental sanity data

Finding appropriate caries prediction models is a recurring subject in odontology, see the review of [14] and the meta-analysis of [8] concerning risk factors for dental caries in young children. A commonly accepted quantification of caries incidence for an individual is the DMFS index, i.e. the count of decayed, missing and filled teeth surfaces. We have at our disposal a data set for 768 French children aged 12 years. We use the variables DMFS index, weight, height, school type (public vs. private), sugar consumption (low vs. high) and consumption of sweetened drinks (low vs. high). An important quantity with respect to an individual's health is the so-called *Body Mass Index* BMI. It is obtained by dividing the weight (in kg) by the squared height (in m), thus defining a continuous scale ranging from underweight to obesity (cf. [6]). A discussion of the association between caries counts and BMI in a more traditional statistical context is given in [15]. There the question was raised if non-linear regression techniques could better identify a possible dependence on BMI, since no definite conclusion on significance of BMI could be drawn. Figure 2 shows a scatterplot of DMFS and BMI data, including a kernel density estimate of BMI and a mobile mean of DMFS with respect to BMI. The mobile mean is calculated over a neighborhood of 25 BMI values at each side. Naively, one might expect either no effect or e.g. a monotonic increase of mean DMFS values with increasing BMI. Yet the mobile mean shows two peaks and a dent in-between in a region of high density of BMI observations, with an overall slightly increasing trend. Moreover, zero observations are abundant in DMFS values. The ZI spline regression models are therefore appropriate candidates for a formal modelisation of these data. We fitted a wide variety of models determined by all combinations of the following options:

- with vs. without ZI
- BMI: linear in the count or the structural zero component or in both
  - regression constant only vs. regression constant and coefficient



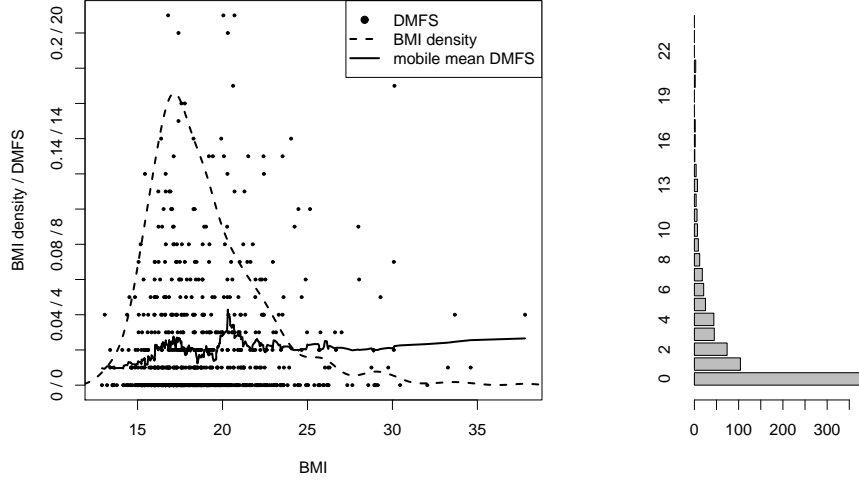


Figure 2: Left: Scatterplot of DMFS and BMI data. Right: DMFS frequencies.

- with vs. without a coefficient for school type
- with vs. without a coefficient for sugar consumption
- with vs. without a coefficient for consumption of sweetened drinks
- spline regression
  - adaptive knots vs. fixed equiprobable quantile knots
  - linear vs. cubic vs. natural cubic splines

For spline curves, we considered at least one knot and a maximum number of knots such that the optimal AIC values could be identified. Since we expect non-linear effects to be rather weak, we neglect BIC in accordance with the results of the preceding simulation study. We restrain the following analysis to finding a single model that fits well. Depending on what we want to know about the process that generates the data, subsets of the considered model classes may be of interest – e.g. linear variable knot models to identify breakpoints. We summarize some noticeable characteristics of the fitted models:

Negative binomial models are clearly superior to Poisson models. So even after including ZI into the model, there remains overdispersion in the count component.

A linear spline model with ZI and 3 fixed interior knots takes the top position with respect to AIC.

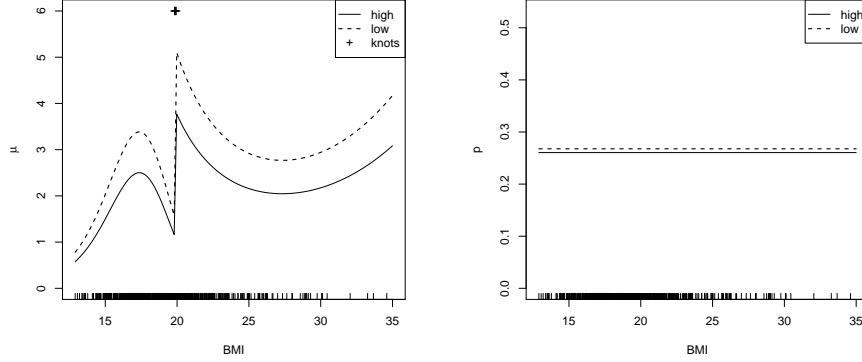


Figure 3: Winner model. Includes the factor for the consumption of sweetened drinks and 4 (almost coinciding) knots. The rug indicates the observed DMFS data.

The factor covariates turn up in the top AIC ranks. In particular, the consumption of sweetened drinks is included in the two top models.

We mention that the strong penalization of BIC favors simple linear models without ZI. In fact, the nonconditional model with no influence of the covariates turns out best.

We found that mean residual errors with leave-one-out cross-validation (MRE) for the top 20 AIC models vary little between 2.410 and 2.433. We look closer at the top AIC model and the best model with respect to MRE which we choose as "winner", see Figures 3 and 4 for predictions of structural zero probability and mean count  $\mu$ . The zero probabilities do not depend on BMI. In both models, the factor for high consumption of sweetened drinks is included and increases the count expectation clearly. Curiously, the probability of structural zeros increases slightly with high consumption. We conjecture that actually there is no effect of this factor on structural zeros, i.e., that the corresponding coefficient is 0. The winner model is the only variable knot model among the 20 preselected models. The plot for the structural zero probability is omitted since it is virtually identical to the one in Figure 4. The model applies natural splines and 4 knots that almost coincide, with  $\varepsilon = 0.001 \times \text{range}(\text{BMI})$  as minimal knot distance in the EBOK algorithm. There is an abrupt jump in the predicted values around the interior knot locations. In summary, both models reproduce the two peaks and a dent in-between, as revealed by the application of a mobile mean. The winner model's abrupt jump in the dent is even more baffling. An explanation for this interesting finding has yet to be established by odontological researchers.

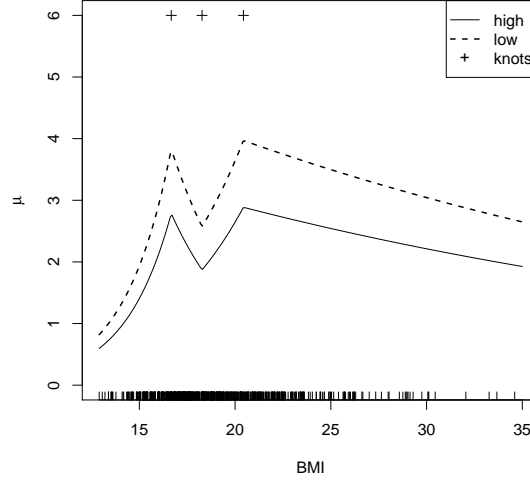


Figure 4: Best AIC model. With factor for consumption of sweetened drinks and 3 (fixed) knots. The rug indicates the observed DMFS data.

## 7 Conclusion and outlook

We presented a class of models that provides high flexibility for regression analysis of count data. In particular, we allow for a binary effect that decides on membership to structural zero observations. The spline approach serves to identify non-linear relations. This is of great importance in fields like medicine or disease control in order to correctly assess the influence of risk factors. When computational complexity of variable knots is manageable, the goodness of the spline fit tends to improve even with less knots than in the fixed knot scenario. A refinement of model choice may prevent a biased selection of the winner model which is due to the usually large number of eligible models. In general, multiple tests and the comparison of different goodness-of-fit criteria help improve the model selection procedure. Criteria for high-dimensional model spaces like the slope heuristic could be considered. Since any parametric linear model is equivalent to a semi-parametric spline model, the decision if a spline model fits better could be based on procedures operating only on spline model fits. Then the pooling of parametric and semi-parametric models, as done in the presented application, can be avoided for model selection. We showed the utility of such model fitting in an application to dental sanity data where classical GLM models and ZI-GLM models failed to reveal a significant influence of the Body-Mass-Index on caries counts.

## References

- [1] D. Boehning, E. Dietz, P. Schlattmann, L. Mendonça, and U. Kirchner. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society: Series A*, 162(2):195–209, 1999.
- [2] Y. B. Cheung. Zero-inflated models for regression analysis of count data: a study of growth and development. *Statistics in medicine*, 21(10):1461–1469, 2002.
- [3] C. Czado, V. Erhardt, A. Min, and S. Wagner. Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. *Statistical Modelling*, 7(2):125–153, 2007.
- [4] C. de Boor. *A Practical Guide to Splines*. Springer, New York, revised edition, 2001.
- [5] R. L. Eubank. *Spline smoothing and nonparametric regression*. M. Dekker, New York, 1988.
- [6] L. A. Ferrara, editor. *Focus on Body Mass Index and Health Research*. Nova Science (New York), first edition, 2006.
- [7] J. H. Friedman. Multivariate Adaptive Regression Splines. *Annals of Statistics*, 19:1–67, 1991.
- [8] R. Harris, A.D. Nicoll, P.M. Adair, and C.M. Pine. Risk factors for dental caries in young children: a systematic review of the literature. *Community Dental Health*, 21:71–85, 2004.
- [9] T. J. Hastie and R. J. Tibshirani. *Generalized additive models*. Chapman & Hall/CRC, 1990.
- [10] D. L. B. Jupp. The lethargy theorem, a property of approximation by  $\gamma$ -polynomials. *Journal of Approximation Theory*, 14:204–217, 1975.
- [11] T. Khoshgoftaar, K. Gao, and R. M. Szabo. An application of zero-inflated poisson regression for software fault prediction. In *Proceedings of the 12th International Symposium on Software Reliability Engineering, 2001*, pages 66–73. Published by the IEEE Computer Society, 2001.
- [12] D. Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.
- [13] N. Molinari, J. F. Durand, and R. Sabatier. Bounded optimal knots for regression splines. *Computational Statistics & Data Analysis*, 45:159–178, 2004.

- [14] L.V. Powell. Caries prediction: A review of the literature. *Community Dentistry and Oral Epidemiology*, 26(6):361–371, 2007.
- [15] P. Tramini, N. Molinari, M. Tentscher, C. Demattei, and A. G. Schulte. Association between caries experience and Body Mass Index in 12-year-old french children. *Caries Research*, 43(6):468–473, 2010.

Table 1: Study (1): Linear fit and cubic spline fits for 100 simulations.

$\alpha = 0.5$		lin.	1	2	3
$\ \cdot\ _\infty$	best	<b>37</b>	25	16	22
	med.	0.4787	0.4796	<b>0.4633</b>	0.4653
	sd	(0.2536)	(0.3781)	(0.4012)	(0.4441)
$\ \cdot\ _1$	best	<b>89</b>	5	5	1
	med.	<b>0.1928</b>	0.2971	0.3317	0.364
	sd	(0.1147)	(0.1243)	(0.126)	(0.1495)
MRE	best	<b>35</b>	28	21	16
	med.	0.5871	0.5801	0.5809	<b>0.5787</b>
	sd	(0.07391)	(0.07362)	(0.07337)	(0.07315)
AIC	best	<b>71</b>	16	6	7
	med.	<b>330.3</b>	330.5	331.9	331.3
	sd	(32.16)	(31.72)	(31.83)	(31.93)
BIC	best	<b>100</b>	0	0	0
	med.	<b>343.5</b>	353.6	358.3	361
$\beta_0^z$	med.	0.9773	0.8133	0.7484	0.7664
	sd	(0.6067)	(0.6447)	(0.6587)	(0.6576)
$\beta_1^z$	med.	-0.8938	-0.7376	-0.7058	-0.7598
	sd	(1.098)	(1.667)	(1.688)	(1.658)

Table 2: Study (1): Linear fit and cubic spline fits for 100 simulations.

$\alpha = 1$		lin.	1	2	3
$\ \cdot\ _\infty$	best	12	15	<b>48</b>	25
	med.	1.001	0.8161	<b>0.5528</b>	0.6432
	sd	(0.3142)	(0.451)	(0.4037)	(0.6481)
$\ \cdot\ _1$	best	<b>64</b>	9	20	7
	med.	<b>0.2542</b>	0.3256	0.3409	0.3907
	sd	(0.1047)	(0.1157)	(0.1447)	(0.1531)
MRE	best	25	20	<b>35</b>	20
	med.	0.5982	0.5989	<b>0.5921</b>	0.5956
	sd	(0.06855)	(0.06817)	(0.06734)	(0.06885)
AIC	best	<b>46</b>	12	28	14
	med.	329.7	329.6	<b>328.1</b>	329.3
	sd	(29.71)	(29.41)	(29.32)	(28.99)
BIC	best	<b>98</b>	0	2	0
	med.	<b>342.9</b>	352.7	354.5	358.9
$\beta_0^z$	med.	0.9504	0.8906	0.771	0.7398
	sd	(0.6368)	(0.5333)	(0.5237)	(0.5268)
$\beta_1^z$	med.	-0.7811	-0.6593	-0.6756	-0.6786
	sd	(1.279)	(1.092)	(1.079)	(1.096)

Table 3: Study (1): Linear fit and cubic spline fits for 100 simulations.

$\alpha = 2$		lin.	1	2	3
$\ \cdot\ _\infty$	best	2	2	<b>59</b>	37
	med.	1.721	1.454	<b>0.5498</b>	0.6712
	sd	(0.4325)	(0.7486)	(0.4782)	(0.6693)
$\ \cdot\ _1$	best	16	6	<b>61</b>	17
	med.	0.4552	0.4504	<b>0.347</b>	0.3867
	sd	(0.08129)	(0.1226)	(0.1642)	(0.1893)
MRE	best	22	7	<b>38</b>	33
	med.	0.6035	0.6115	<b>0.5969</b>	0.6
	sd	(0.08173)	(0.07994)	(0.07858)	(127.2)
AIC	best	12	6	<b>59</b>	23
	med.	323.3	322.1	316.6	<b>316.6</b>
	sd	(32.37)	(30.98)	(29.35)	(29.68)
BIC	best	<b>73</b>	2	25	0
	med.	<b>336.5</b>	345.2	343	346.3
$\beta_0^z$	med.	1.255	1.08	0.9983	0.972
	sd	(0.5357)	(1.001)	(0.523)	(0.5195)
$\beta_1^z$	med.	-1.269	-0.952	-1.023	-1.057
	sd	(1.122)	(1.46)	(1.126)	(1.108)

Table 4: Study (1): Linear fit and cubic spline fits for 100 simulations.

$\alpha = 3$		lin.	1	2	3
$\ \cdot\ _\infty$	best	0	0	<b>63</b>	37
	med.	2.525	2.569	0.5505	<b>0.5438</b>
	sd	(0.4793)	(1.237)	(0.4578)	(0.3987)
$\ \cdot\ _1$	best	2	1	<b>85</b>	12
	med.	0.7175	0.6292	<b>0.2804</b>	0.3449
	sd	(0.1066)	(0.1043)	(0.162)	(0.1786)
MRE	best	20	4	<b>39</b>	37
	med.	0.7665	0.7564	0.7341	<b>0.7323</b>
	sd	(0.1283)	(0.1868)	(0.1112)	(0.1171)
AIC	best	2	1	<b>71</b>	26
	med.	363.3	352.9	<b>335.8</b>	338.6
	sd	(42.61)	(35.84)	(31.55)	(31.56)
BIC	best	19	1	<b>77</b>	3
	med.	376.5	376	<b>362.2</b>	368.3
$\beta_0^z$	med.	1.519	1.083	0.9298	0.8992
	sd	(0.6304)	(0.5472)	(0.505)	(0.4898)
$\beta_1^z$	med.	-1.662	-0.8224	-1.091	-1.034
	sd	(2.16)	(1.358)	(1.017)	(0.9697)

Table 5: Study (2): Linear fit and linear spline fits for 50 simulations.

fixed knots		lin.	1	2	3	4	5	6
$\ \cdot\ _\infty$	best	0	0	0	0	2	12	<b>36</b>
	med.	1.321	1.429	1.303	1.254	0.7911	0.6064	<b>0.4843</b>
	sd	(0.08845)	(0.1048)	(0.2739)	(0.3453)	(0.3474)	(0.2732)	(0.2471)
$\ \cdot\ _1$	best	0	0	0	0	1	13	<b>36</b>
	med.	0.5916	0.588	0.4133	0.5209	0.2603	0.1994	<b>0.1736</b>
	sd	(0.02193)	(0.02282)	(0.03472)	(0.03662)	(0.03817)	(0.0459)	(0.03317)
MRE	best	5	1	6	1	10	<b>14</b>	13
	med.	3.23	3.225	3.099	3.171	3.104	3.083	<b>3.083</b>
	sd	(0.3565)	(0.3487)	(0.2872)	(0.3025)	(0.2665)	(0.2525)	(0.2414)
AIC	best	0	0	0	0	4	16	<b>30</b>
	med.	760.5	756.9	660.8	697.6	616.2	<b>600.7</b>	602.2
	sd	(69.02)	(66.07)	(52.34)	(49.92)	(45.74)	(40.39)	(39.84)
BIC	best	0	0	0	0	11	<b>20</b>	19
	med.	773.6	773.4	680.6	720.7	642.6	<b>630.4</b>	635.1

Table 6: Study (2): Linear fit and linear spline fits for 50 simulations.

var. knots		1	2	3	4	5	6
$\ \cdot\ _\infty$	best	0	0	1	16	15	<b>18</b>
	med.	1.387	1.286	1.14	0.4642	<b>0.442</b>	0.4773
	sd	(0.08547)	(0.3015)	(0.3556)	(0.2394)	(0.2146)	(0.2337)
$\ \cdot\ _1$	best	0	0	0	<b>24</b>	15	11
	med.	0.5387	0.3356	0.2626	<b>0.163</b>	0.1871	0.1849
	sd	(0.01319)	(0.02739)	(0.06051)	(0.04077)	(0.04256)	(0.05947)
MRE	best	10	6	<b>12</b>	9	3	10
	med.	3.189	3.104	3.12	3.118	3.099	<b>3.086</b>
	sd	(0.3405)	(0.2687)	(0.248)	(0.2352)	(0.2144)	(0.2611)
AIC	best	1	3	5	<b>18</b>	14	9
	med.	703.6	639.7	621.5	<b>593</b>	593.5	598.3
	sd	(57.8)	(45.73)	(43)	(39.36)	(35.93)	(39.84)
BIC	best	2	3	8	<b>21</b>	12	4
	med.	723.4	666.1	654.5	<b>632.6</b>	639.6	651



Table 7: Study (2): Cubic spline fits for 50 simulations.

fixed knots		1	2	3	4	5	6
$\ \cdot\ _\infty$	best	0	<b>15</b>	1	14	11	9
	med.	1.309	0.3749	0.5663	<b>0.3207</b>	0.3449	0.3494
	sd	(0.7008)	(0.2088)	(0.35)	(0.2479)	(0.3341)	(0.3235)
$\ \cdot\ _1$	best	0	11	0	14	<b>15</b>	10
	med.	0.5636	0.1609	0.2504	<b>0.1347</b>	0.1353	0.1446
	sd	(0.02085)	(0.03708)	(0.0466)	(0.04189)	(0.0518)	(0.05427)
MRE	best	8	<b>14</b>	6	7	9	6
	med.	3.238	3.103	3.154	3.101	<b>3.088</b>	3.108
	sd	(0.3993)	(0.2511)	(0.2568)	(0.2449)	(0.2374)	(0.2364)
AIC	best	0	12	0	<b>23</b>	11	4
	med.	714.2	598.5	616.2	595.9	<b>595.9</b>	599.2
	sd	(57.14)	(42.12)	(39.69)	(39.18)	(38.6)	(38.8)
BIC	best	0	<b>37</b>	1	11	1	0
	med.	737.3	<b>624.8</b>	645.9	628.9	632.2	638.8

Table 8: Study (2): Cubic spline fits for 50 simulations.

var. knots		1	2	3	4	5	6
$\ \cdot\ _\infty$	best	0	<b>29</b>	8	4	4	5
	med.	1.357	<b>0.2841</b>	0.348	0.4517	0.4851	0.6122
	sd	(0.7506)	(0.1916)	(1.514)	(1.926)	(2.918)	(1.886)
$\ \cdot\ _1$	best	0	<b>32</b>	10	3	4	1
	med.	0.4675	<b>0.126</b>	0.1382	0.1713	0.1883	0.2139
	sd	(0.02244)	(0.04604)	(0.4086)	(0.2071)	(0.13)	(0.5814)
MRE	best	6	<b>14</b>	5	7	7	11
	med.	3.202	<b>3.107</b>	3.112	3.135	3.126	3.121
	sd	(0.3382)	(0.2447)	(0.4305)	(2.431)	(28.84)	(2285)
AIC	best	0	<b>31</b>	7	3	1	8
	med.	692	<b>598.6</b>	599.9	601.8	602.1	601.3
	sd	(50.82)	(39.37)	(39.02)	(36.1)	(39.31)	(39.43)
BIC	best	0	<b>41</b>	1	1	0	7
	med.	718.4	<b>631.6</b>	639.5	647.9	654.9	660.7